Multimed Tools Appl
DOI 10.1007/s11042-014-2261-2

Land cover classification using geo-referenced photos

Daniel Leung · Shawn Newsam

Received: 21 February 2014 / Revised: 9 July 2014 / Accepted: 3 September 2014

Abstract We investigate publicly available geo-referenced photo collections for land cover classification. Mapping land cover is a fundamental task in the geographic sciences and is typically done using remote sensing (overhead) imagery through manual annotation. We here propose a novel alternate approach based on proximate sensing. The goal of proximate sensing is to map what-is-where on the surface of the Earth using ground level images of objects and scenes. It has the potential to map phenomena not observable through remote sensing. We perform an extensive case study on using ground level images for binary land cover classification into developed and undeveloped regions. We investigate visual features and text annotations to label images or sets of images with these two classes. Knowing the location of the images allows us to generate land cover maps which we quantitatively evaluate using ground truth maps. We apply our approach to two photo collections, Flickr, the popular photo sharing website, and the Geograph project, whose goal is to collect geographically informative photos. Comparing these two collections allows us to measure the impact of photographer intent. We utilize a weakly supervised learning framework which eliminates the need for manually labeled training data. We also investigate methods for filtering images that are unlikely to be geographically informative. Our results are promising and validate proximate sensing as a novel alternate approach to geographic discovery.

Keywords Proximate sensing · Land cover classification · Geo-referenced photos

1 Introduction

Remote sensing uses overhead images, such as acquired from air- or space-borne platforms, to map what-is-where on the surface of the Earth. We have established an equivalent novel

D. Leung (✉) · S. Newsam
Electrical Engineering and Computer Science, University of California, Merced, CA 95343 USA
e-mail: cleung3@ucmerced.edu

S. Newsam
e-mail: snewsam@ucmerced.edu

Published online: 19 September 2014

framework termed *proximate sensing* [17, 18] which instead uses ground level images of objects or scenes. Proximate sensing is made possible by the confluence of digital cameras and global positioning systems (GPS) (or similar mechanisms for assigning location to an image) and has the potential to map phenomena not observable from above. It is also made possible and provides interesting research challenges for many areas of multimedia content analysis such as automated image understanding.

The popularity and rapid growth of photo sharing websites and other sources of user contributed geo-referenced[1] images enables proximate sensing to be applied on a scale rivaling that of remote sensing. At present, Flickr[2] has over 200 million geotagged items and there are hundreds of millions of geo-referenced images available at other photo sharing websites such as Panoramio,[3] Picasa,[4] and KanKan;[5] travel photography websites such as TrekEarth;[6] nature photography websites such as TrekNature;[7] user-maintained encyclopedic websites such as Wikipedia;[8] and even individual projects such as that of Tom Graham who walked every street of San Francisco, capturing thousands of photographs which he is making available on-line along with other travel logs.[9] The interesting challenge to the multimedia community is how to use this largely unstructured data that is being acquired by millions of so-called citizen sensors to perform geographic knowledge discovery.

The recent phenomenon of volunteered geographic information (VGI) provides a broader context for geographic discovery using publicly available photo collections. Geographer Michael Goodchild coined the term VGI in 2007 [8] to refer to the growing collections of geographically relevant information provided voluntarily by individuals. Enabled by emerging technologies centered around the web, VGI is creating sources of geographic information that differ along many dimensions from traditional ones. While some of these differences present challenges, such as the legitimacy of the contributors and the relative lack of provenance information, others enable large-scale geographic knowledge discovery not possible before in terms of reduced temporal latency and providing the “people’s” perspective. In a recent position paper [23], we postulated that many of the techniques recently developed for leveraging geo-referenced community contributed photos to perform automated annotation and other tasks are really a form of VGI, albeit often a serendipitous one.

The main contribution of this paper is a novel framework which uses state-of-the-art techniques in multimedia content analysis, in particular automated image understanding and statistical text analysis, to perform geographic knowledge discovery in large collections of on-line photos. We focus on land cover classification and present an extensive case study

[1] We use the term geo-referenced to indicate that a multimedia object has at least approximate location metadata associated with it.

[2] http://www.flickr.com/

[3] http://www.panoramio.com/

[4] http://picasaweb.google.com

[5] http://www.kankanchina.cn/

[6] http://www.trekearth.com/

[7] http://www.treknature.com/

[8] http://www.wikipedia.org/

[9] http://sfwalkingman.com/

in which images from Flickr and the Geograph British Isles project[10] are used to classify a large region of the United Kingdom into develop and undeveloped regions.

The salient aspects of our work include: 1) we compare different visual features and text annotations for classifying individual or sets of images as depicting a developed or undeveloped region; 2) we evaluate the effect of photographer intent by comparing the two different photo collections; 3) we propose a weakly supervised learning framework which eliminates the need for manually labelled ground truth data; 4) we perform quantitative evaluation using ground truth maps; and 5) we investigate methods for filtering images that are unlikely to be geographically informative.

The remainder of the paper is as follows. First, we describe related work on geo-referenced photo collections and situate our work in the class of problems which focus on geographic discovery. We then describe our land cover classification framework and present experimental results. This is followed by an investigation into removing uninformative images. Finally, we finish with some thoughts on future research directions.

2 Related work

There is a growing body of research on analyzing geo-referenced community contributed photo collections. Methods have been developed which leverage the collections to 1) annotate novel images; 2) annotate geographic locations; and 3) perform geographic knowledge discovery. Our work on proximate sensing for land cover classification is an example of this last class.

2.1 Leveraging collections to annotate novel images

Automated annotation is essential for managing large image collections. Methods have been developed that leverage large sets of geo-referenced images to semantically annotate novel images whose location is known. This is particularly useful for images captured using GPS enabled cameras or camera phones as the system generated annotation allows the images to be organized and searched at a more meaningful way than with low-level image descriptors such as color or texture. Methods have been developed for suggesting tags such as “surfer”, “wave”, and “Santa Barbara” for a photograph of someone surfing in Santa Barbara, California [21]; for assigning a constrained set of event/activity labels such as “a visit to the beach” or “wedding” [14]; for annotating groups of images at the event (“skiing”) or scene (“coast”) level [1]; for annotating the identities of people appearing in an image [22]; and for linking images, such as a photograph of the Arc de Triomphe, to relevant Wikipedia articles [25].

Collections of geo-referenced images have also been used to annotate the locations of novel images–that is, to estimate where in the world the photo was taken. Methods have been developed to geo-locate web cameras distributed around the United States based on image variations relating to the diurnal cycle and weather [13]; to geo-locate a single image using only its visual content [10] as well as textual tags [7] by performing similarity search against a reference collection; and to estimate coarse image location by first clustering a reference

[10] http://www.geograph.org.uk/

collection and then indexing the novel image based on its visual content and textual tags [2, 5, 6].

2.2 Leveraging collections to annotate geographic locations

Collections of geo-referenced images have also been used to annotate geographic locations, a task in which on-line photo collections are considered more explicitly as VGI as the objective is more in line with the problem of determining what-is-where on the surface of the Earth. Methods have been developed for visually annotating prominent landmarks with representative images at the city [5] and world-wide [30] scales; to suggest representative tags as well as images for geographic locations [15, 16, 22]; and to automatically generate tourist maps showing popular landmarks as vectorized icons [4].

2.3 Leveraging collections for geographic knowledge discovery

The goal of geographic knowledge discovery is to learn what-is-where on the surface of the Earth in the broad sense of the term “what”. It seeks to generate maps not only of the physical aspects of our world, such as the terrain, but also of the abstract aspects, such as culture and natural or man-made behavior. The growing availability and unique perspective of geo-referenced ground level images and videos coupled with advances in multimedia content analysis, make proximate sensing a promising alternate to traditional methods for geographic knowledge discovery.

There has been relatively little work on using geo-referenced photos for geographic knowledge discovery. Examples of work in this area include using large collections of geo-referenced images to discover spatially varying (visual) cultural differences among concepts such as “wedding cake” [29]; to discover interesting properties about popular cities and landmarks such as the most photographed locations [5]; to estimate weather satellite images using widely distributed Web cameras [13]; and to create a map-like partitioning of a country-sized region into geographically coherent subregions [6].

Our work in this paper represents a more structured approach to geographic knowledge discovery from geo-referenced photo collections. We present a proof-of-concept case study on land cover classification; however, our framework is general enough to map a wide variety of other phenomena as long as the signals of interest are detectable in the images. Extending the framework beyond land cover classification will be the focus of future work.

3 Case study: land cover classification

This section describes a case study in which geo-referenced photo collections are used to perform land cover classification into developed and undeveloped regions. We focus on this problem to validate proximate sensing as an alternate to traditional methods for geographic knowledge discovery. This is also a problem for which we have ground truth for evaluation.

Figure 1 shows the workflow of our approach. The input is a large collection of geo-referenced ground level images. We perform feature extraction at the image or tile level. Training images or tiles are labelled in a weakly supervised manner using a ground truth map. Binary classifiers are learned and applied to label images or tiles as being developed or undeveloped. These labels are finally used to generate maps which are compared with the ground truth.

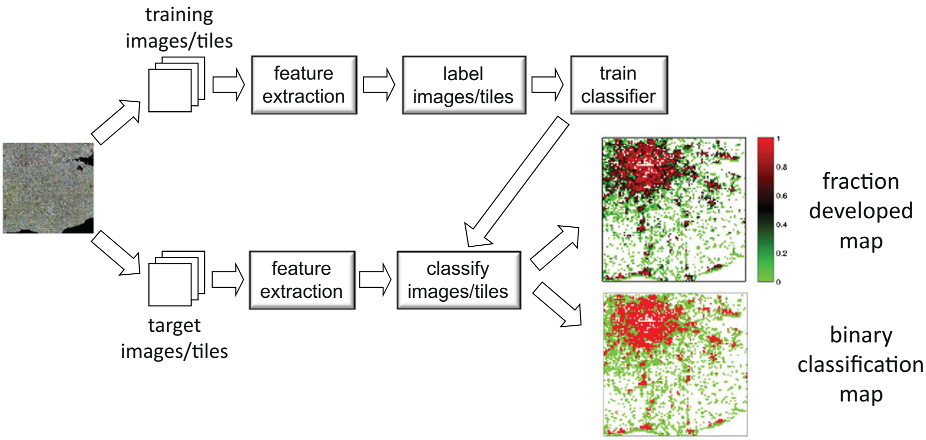


Fig. 1 The workflow of our approach. The input is a large collection of geo-referenced ground level images. The output is a map of developed and undeveloped regions

3.1 Dataset

Our study area is the 100x100 km of Great Britain corresponding to the TQ square in the British national grid system. This region encompasses the London metropolitan area and thus includes developed and undeveloped regions.

We use the publicly accessible Countryside Information System (CIS) to download the Land Cover Map 2000 (LCM2000) of the United Kingdom's Centre for Ecology & Hydrology for the TQ study region. This serves as our ground truth. We focus on the LCM2000 Aggregate Class (AC) data which provides the percentages of ten land cover classes at the 1x1 km scale. Figure 2 shows the dominant classes for the TQ region.

We focus on binary classification into developed and undeveloped regions so we further aggregate the ten land cover classes into a developed superclass consisting of LCM AC 7: Built up areas and gardens, and an undeveloped superclass consisting of the remaining nine

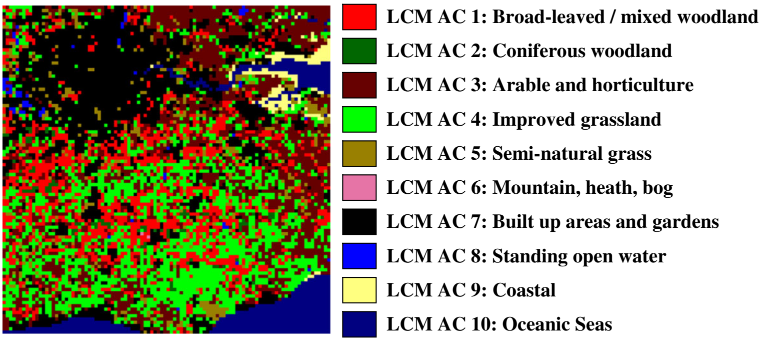

LCM AC 1: Broad-leaved / mixed woodland
LCM AC 2: Coniferous woodland
LCM AC 3: Arable and horticulture
LCM AC 4: Improved grassland
LCM AC 5: Semi-natural grass
LCM AC 6: Mountain, heath, bog
LCM AC 7: Built up areas and gardens
LCM AC 8: Standing open water
LCM AC 9: Coastal
LCM AC 10: Oceanic Seas

Fig. 2 The dominant Land Cover Map 2000 Aggregate Classes (AC) for the TQ study area. This area measures 100x100 km and encompasses the London metropolitan area which appears towards the north-west. Shown are the dominant classes for each 1x1km tile. The dataset also includes the percentages of each class

classes. We derive two ground truth datasets, one which indicates the fraction developed for each of the 10K 1x1 km tiles in the TQ region and another which simply indicates a binary label for each tile by applying a threshold of 0.5 to the fraction developed. We refer to the first dataset as the ground truth *fraction values* and the second as the ground truth *binary labels*. Figure 3 shows the ground truth datasets as maps.

We compile two geo-referenced image collections for the study area. First, we use the Flickr application programming interface (API) to download approximately 920K Flickr images located within the TQ region. The longitude and latitude information provided by the Flickr API is then used to assign each image to a 1x1 km tile. Figure 4a shows the distribution of the Flickr images. While Flickr contains a large collection of geo-referenced images, its spatial coverage is not uniform. For our study area, 5,420 of the 10K 1x1 km tiles do not contain any Flickr images. The 4,580 tiles with images contain an average of 200.7, a median of 10, and a maximum of 53,840 images.

We download a second set of potentially more geographically informative images from the Geograph Britain and Ireland (GBI) project whose aim is to "collect geographically representative photographs and information for every square kilometre of Great Britain and Ireland". This project contains over three million photos contributed by over 10K users and allows us to investigate the effect of photographer intent. We use the GBI API to download approximately 120K Geograph images for the study area. While there are fewer Geograph images, they are more uniformly distributed than the Flickr images as shown in Fig. 4b. Now, only 614 of the 10K tiles do not contain any images and all but a few of these correspond to ocean. The remaining 9,386 tiles contain an average of 12.9, a median of 5, and a maximum of 1,458 images.

In order to investigate whether the Flickr or Geograph images are better for binary land cover classification, we use a *common evaluation dataset consisting of the 4,441 tiles which contain images from both datasets*. These tiles contain over 900K Flickr and over 90K Geograph images. This evaluation dataset is split into disjoint training and test sets with 400 and 4,041 tiles respectively.

Figure 5 shows sample images from the Flickr and Geograph datasets. Pairs of odd/even rows show Flickr/Geograph images for the same 1x1km tiles. The top two pairs of rows are for tiles with a developed fraction of 1.0 while the bottom two pairs of rows are for tiles with

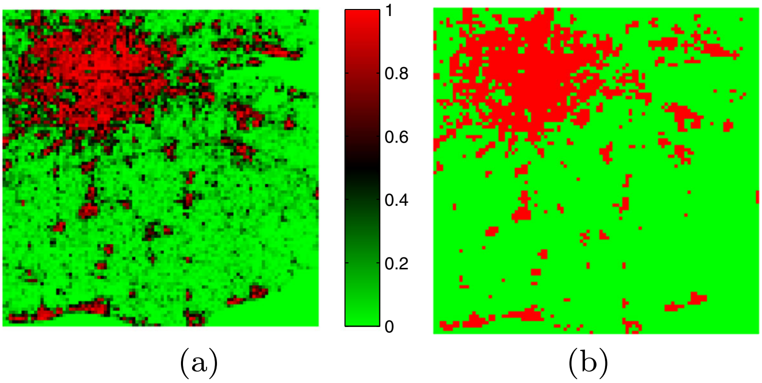


Fig. 3 Ground truth derived from the LCM 2000 AC data. **a** Map of fraction developed values for each 1x1 km tile. **b** Map of binary labels in which *red* and *green* are used to indicate developed and undeveloped tiles respectively. The binary labels are derived from the fraction values by applying a threshold of 0.5

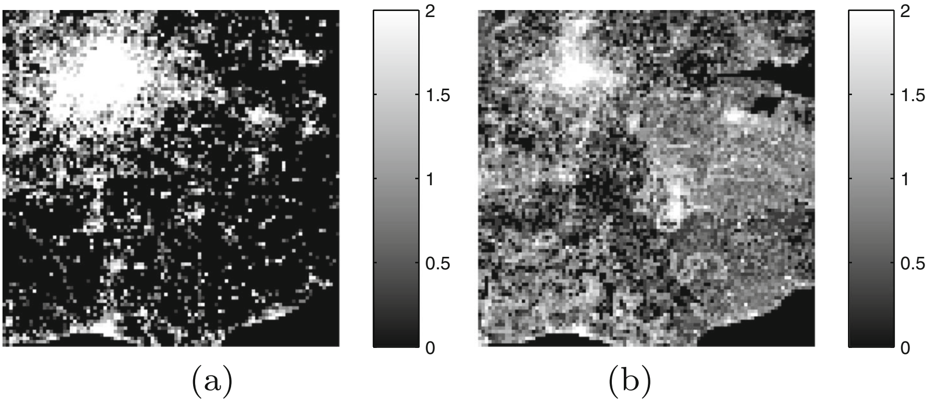


Fig. 4 The distribution of images for the TQ study region in the **a** Flickr and **b** Geograph datasets. On a base-10 logarithmic scale

a developed fraction of 0. The Geograph images tend to be more geographically informative although both support land cover classification as demonstrated in the experiments below.

3.2 Features

We consider three different visual features as well as text features derived from annotations associated with the images.

3.2.1 Color histogram

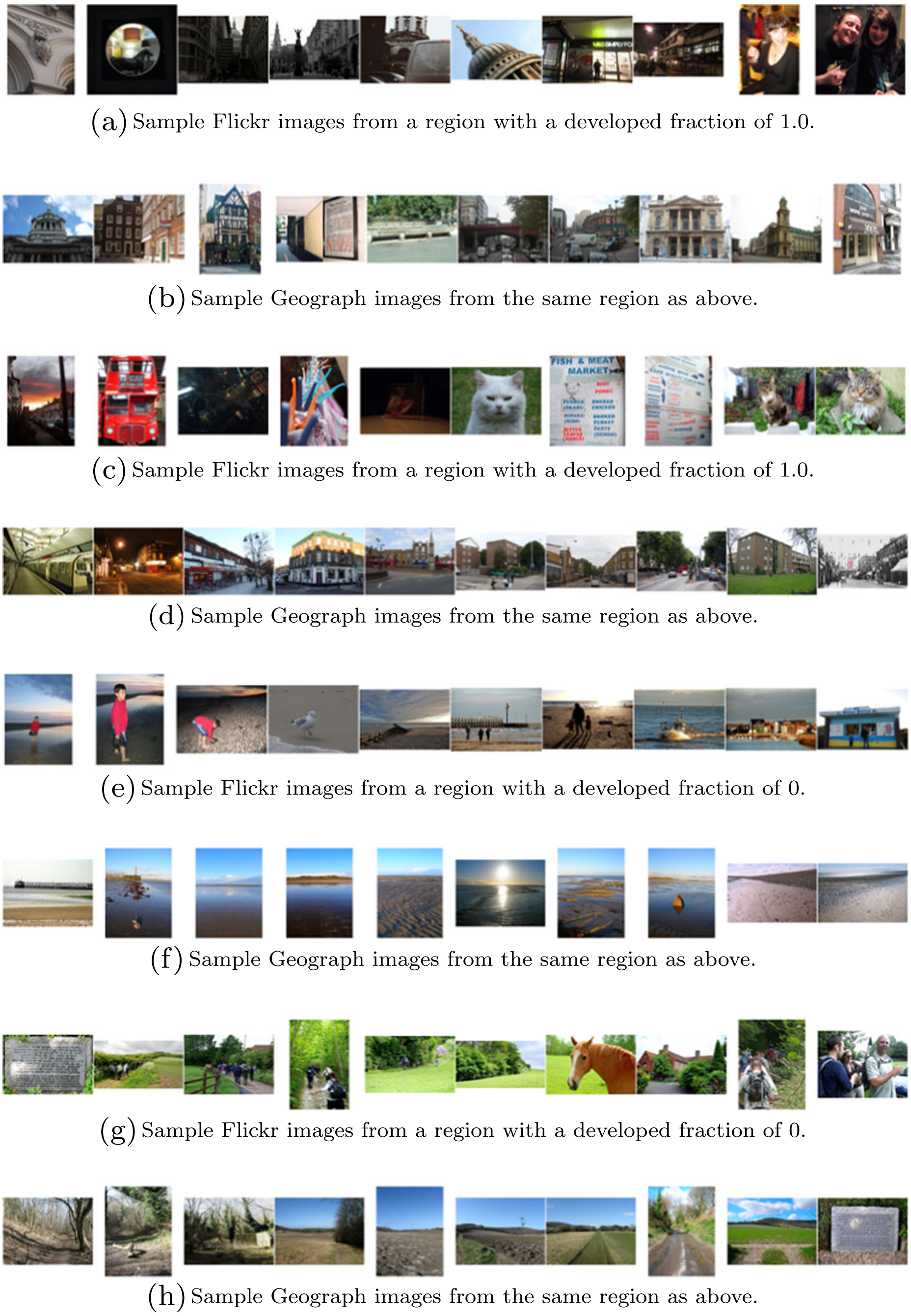
In order to investigate whether color is a discriminating feature for our two-class problem, we extract color histogram descriptors from each image. We transform the images to the hue-lightness-saturation (HLS) colorspace and quantize each dimension into 4 bins for a total feature vector length of 64. The histograms are normalized to have an L1 norm of one to account for different image sizes.

3.2.2 Edge histogram

We extract edge histogram descriptors which quantify the distribution of edges at different orientations. This is motivated by the observation that images of developed scenes typically have a higher proportion of horizontal and vertical edges than images of undeveloped scenes. This is evident in the sample images in Fig. 5. Following the method outlined in [20], we apply a set of five 2x2 linear filters to detect edges at roughly horizontal, vertical, 45° diagonal, 135° diagonal, and isotropic (non-orientation specific) directions. A threshold is applied to the outputs of these filters and the proportions of edges in the different directions are summarized in a five bin L1 normalized histogram.

3.2.3 Gist

The final visual features we consider are Oliva and Torralba's gist descriptors [24] which have shown to be effective at recognizing real world scenes such as coast, mountain, forest, etc. Gist features bypass the segmentation and processing of individual objects or regions,



(a) Sample Flickr images from a region with a developed fraction of 1.0.

(b) Sample Geograph images from the same region as above.

(c) Sample Flickr images from a region with a developed fraction of 1.0.

(d) Sample Geograph images from the same region as above.

(e) Sample Flickr images from a region with a developed fraction of 0.

(f) Sample Geograph images from the same region as above.

(g) Sample Flickr images from a region with a developed fraction of 0.

(h) Sample Geograph images from the same region as above.

Fig. 5 Sample images from the Flickr and Geograph datasets

similar to the histogram features above, and characterize the "spatial envelop" of the scene using frequency-domain techniques. Gist features are similar to texture features extracted using Gabor filters in that they characterize the spectral energy of an image using Gaussian shaped filters tuned to different scales and orientations. We extract 60 dimensional gist feature vectors from each image.

The experiments below compare the performance of the three visual features which, to summarize, include: a 64 dimensional color histogram feature, a five dimensional edge histogram feature, and a 60 dimensional gist feature for each image.

3.2.4 Text

Flickr and Geograph images commonly have user-supplied text associated with them. In the case of the Flickr images, this includes the image titles, descriptions, and tags. For example, the left-most image in Fig. 5a is titled "Roosting dragon"; has the description "Or it might be a vampire bat? In Chancery Lane. Originally uploaded for Guess Where London."; and, is tagged with: "gwl, Guess Where London, stucco, dragon, Guessed by Citymuso, 115A, Chancery Lane, WC2, Holborn, Camden, London, England". The Geograph images have titles, descriptions, and categories. The left-most image in Fig. 5b is titled "The Old Bailey, London"; has the comment "The Central Criminal Court, home of justice in England and Wales."; and is categorized as "Building of civic importance". We therefore investigate whether this user-supplied text is effective for land cover classification.

The text analysis is performed at the tile level since there is typically not enough text associated with the individual images for effective classification. Each of the text components associated with an image obtained within each 1x1 km tile region is parsed into a set of terms (words) which are then pooled among terms from other images within the same tile. At the moment, all terms are given equal weight although different weightings based on the relative importance of the components would be an interesting extension.

It is unlikely that classification at the term level would be effective due to the sparse appearance of terms among the dictionary, so we apply a latent semantic approach from text document analysis in which a hidden topic $z \in Z = \{z_1, ..., z_K\}$ is associated with the observed occurrence of a word $w \in W = \{w_1, ..., w_M\}$ in a document (tile) $d \in D = \{d_1, ..., d_N\}$. This latent layer also helps overcome the problems of synonymy and polysemy.

We use a generative probabilistic technique termed probabilistic latent semantic analysis (pLSA) [11, 12] to learn the hidden topics. A pLSA model can be expressed as

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d),$$

where $P(w|d)$ is the observed word distributions over documents.

We use the Expectation Maximization (EM) algorithm to learn the distribution of words over hidden topics. In the E-step, the posterior probabilities for the hidden topics are evaluated:

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')},$$

while in the M-step, the parameters of E-step are estimated based on the result of E-steps:

$$P(w|z) = \frac{\sum_d n(d, w)P(z|d, w)}{\sum_{d,w'} n(d, w')P(z|d, w')},$$

$$P(d|z) = \frac{\sum_w n(d, w)P(z|d, w)}{\sum_{d',w} n(d', w)P(z|d', w)},$$

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z|d,w),\ R \equiv \sum_{d,w} n(d,w).$$

Instead of defining the number of hidden topics as the number of ground truth classes as is often done in pLSA, we use pLSA as a tool to reduce the dimensionality of the term histogram of each tile by computing the distributions of hidden topics over the tiles, $P(z|d)$. The distributions over hidden topics provide an explicit representation that is more robust than the distributions over terms. To evaluate the hidden topic distribution of a novel tile during classification, the EM algorithm is applied with a fixed $P(w|z)$ learned from the training step.

We first determine reasonably sized term-dictionaries for each of the datasets. After applying stopping and stemming, a total of 106,213 unique terms result from the over 900K Flickr images in the 4,441 tiles with Flickr and Geograph images that have text, and 31,056 unique terms result from the over 90K Geograph images from the same tiles. The dictionary for the Flickr dataset is selected as the 2,708 most frequent Flickr terms, and the dictionary for the Geograph dataset is selected as the 2,702 most frequent Geograph terms.

A term histogram is computed for each tile based on the union of terms from all the images in the tile. The histograms for 200 training tiles are combined into a term-document matrix and pLSA is used to learn the term-topic distributions for a 60 topic model (this number was chosen empirically based on performance). Finally, a topic distribution is computed for each of the 4,041 tiles in the test set using the pLSA machinery.

To summarize, each tile is represented with a 60 dimensional text feature vector that consists of the distribution over the latent topics.

3.3 Experiments

The goal of the experiments is to use the geo-referenced images as represented by their visual or text features to produce developed/undeveloped land cover maps. We formulate this as a supervised classification problem in which support vector machines (SVMs) are trained on a labeled subset of the data and are then used to assign labels to a disjoint held-out set. We compare applying the SVMs 1) at the image level, in which case the image labels (developed and undeveloped) are aggregated to produce the final tile level fraction values and binary labels, and 2) applying them directly at the tile level. These two modes are described in Sections 3.3.1 and 3.3.2 below. Performance is evaluated by comparing the predicted maps to the ground truth derived from the Land Cover Map 2000.

The SVMs are implemented using the LIBSVM package [3]. We use radial basis function (RBF) kernels and determine optimal values for the two parameters, the penalty term C and the kernel width γ, through grid-search on a random partitioning of the training set.

3.3.1 Image level classification

In this set of experiments, the SVMs are used to classify individual images as being developed or undeveloped. These labels are then aggregated to determine the tile fraction values and binary labels for comparison with the ground truth.

Training the SVMs requires a set of images labeled as developed or undeveloped. We construct a weakly labeled training set in a completely automated manner by propagating the ground truth labels (developed or undeveloped) of the tiles containing the training images as follows. First, we identify the 100 most developed and the 300 least developed tiles according to the ground truth fraction values. These are training tiles. We use 300 least developed tiles because the least developed tiles generally contain fewer images than the most developed tiles. We then randomly sample approximately 2,500 images from the 100 most developed tiles and label them as developed. We similarly sample and label as undeveloped approximately 2,500 images from the 300 least developed tiles. This results in labeled training sets containing 5,031 and 5,026 images for the Flickr and Geograph datasets respectively.

Such a weakly labeled training set has two important advantages over a manually labeled one. First, it does not require human effort. Second, it avoids the subjective interpretation of what is meant by developed at the image level. Indeed, we showed in previous work [18] that a weakly labeled training set outperforms one in which the labels are assigned manually.

The SVMs are trained using the visual features and labels of the 5K+ training images. They are then used to label each of the images in the 4,041 test tiles that remain after the 400 training tiles have been removed (thus the training and test sets are distinct). These labels are aggregated to compute two tile level values: a fraction developed which is the number of images in the tile labeled as developed by the SVM divided by the total number of test images in the tile; and a binary label which is determined by applying a threshold to the fraction developed. We explore using a threshold fixed at 0.5 as well as an adaptive threshold that is chosen so that the ratio of developed to undeveloped tiles in the predicted set matches that of the ground truth (this represents prior knowledge of the problem in the form of the expected ratio of developed to undeveloped regions).

3.3.2 Tile level classification

In this set of experiments, the SVMs are used to label the tiles directly. A single visual feature is computed for each tile by averaging the features from all the images located in that tile. This has the simple interpretation of a tile level histogram for the edge and color features. For the gist features, it is the average over all the images of the spectral energy in each of the frequency channels corresponding to the Gabor filters. The text features are already computed at the tile level so no aggregation is needed.

The training set is the 100 features (visual or text) corresponding to the 100 most developed tiles and the 100 features corresponding to the 100 least developed tiles. These 200 training tiles are a subset of the 400 training tiles used in the image level classification above.

Once trained, the SVMs are used to label each of the 4,041 test tiles as developed or undeveloped again using a single feature. Note that this results in only a binary label for each tile; the fraction developed value is not estimated when the SVM labeling is done at the tile level (we have not yet considered using the classifier margin for this).

3.4 Results

The results are evaluated by comparing the predicted fraction values and binary labels to that of the ground truth for the 4,041 tiles in the test set. The ground truth fraction values and binary labels for the test set are shown in the maps in Fig. 6.

The predicted fraction values are evaluated based on their correlation with the ground truth values. Let f_i' and f_i represent the predicted and ground truth values for tile i where $i = 1, \ldots, 4041$. We compute the coefficient of correlation as $\rho_{f'f} = cov(f', f)/\sigma_{f'}\sigma_f$ where $cov(f', f)$ is the covariance of f and f' over i and $\sigma_{f'}$ (σ_f) is the standard deviation of f' (f) over i. $\rho_{f'f}$ ranges from -1 to 1 with a value of 0 indicating no correlation, and values of -1 and 1 indicating strong negative and positive correlation respectively.

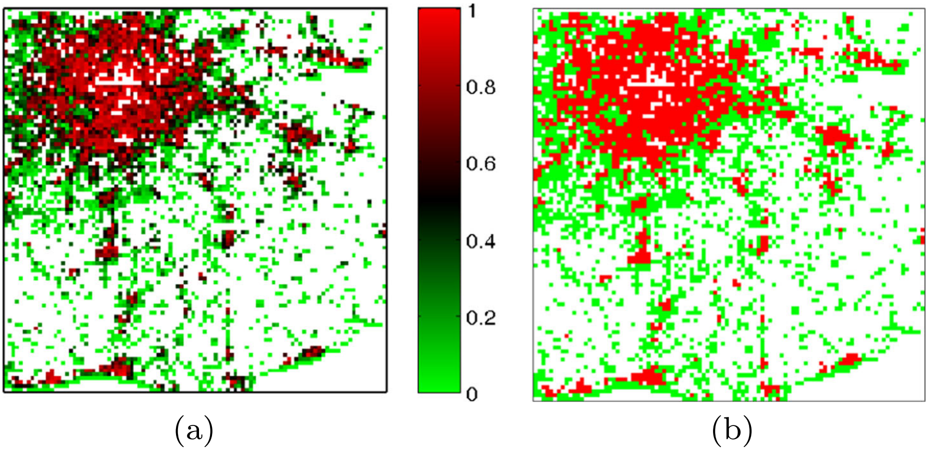


Fig. 6 Ground truth data for the 4,041 tiles in the test set. **a** Fraction map indicating the percent developed for each 1x1 km tile. **b** Binary map indicating the tiles labeled as developed (*red*) or undeveloped (*green*). The *white regions* are where there are no images and are thus not used in the evaluation

The binary labels are evaluated using classification rates. The overall classification rate is the percentage of tiles assigned the same label–developed or undeveloped–as the ground truth.

As mentioned earlier, when the SVM labeling is performed at the image level, the binary labels are determined by applying either a fixed or an adaptive threshold to the fraction values. In the fixed case, a tile is labeled as developed if the fraction value is greater than 0.5 (i.e., more than 50 % of its images are labeled as developed). The adaptive threshold is chosen so that the resulting ratio of developed to undeveloped tiles matches that of the ground truth.

Since 2,386 of the 4,041 test tiles are undeveloped in the ground truth, labeling all images or all tiles as undeveloped results in a "chance" classification rate of 59.0 %.

3.4.1 Results of the image level classification

The third to fifth columns of Table 1 summarize the results when the SVM classification is performed at the image level. The predicted fraction values and binary labels for the best case corresponding to Geograph images classified using gist features are shown visually in Fig. 7. Compare this with the ground truth in Fig. 6.

Based on these results, we conclude the following about the image level classification:

- The Geograph dataset outperforms the Flickr dataset.
- The gist features perform best overall. The edge histogram features perform better than the color histogram features. This ordering is true for both datasets.
- The adaptive threshold improves the overall classification rate in most cases.

3.4.2 Results of the tile level classification

The last column of Table 1 summarizes the results when the SVM classification is performed at the tile level. Based on these results, we conclude the following about the tile level classification:

Table 1 The results of the image and tile level classifications

		Image level classification			Tile level classification
		Binary prediction			
		Overall class. rate			Binary prediction
Dataset	Feature	Fixed Threshold %	Adaptive Threshold %	Fraction Prediction ρ	Overall Class. Rate (%)
Geograph	Color	70.9	73.0	0.519	68.8
Geograph	Edge	70.7	73.1	0.528	72.2
Geograph	Gist	**75.0**	**75.0**	**0.614**	74.0
Geograph	Text	N/A	N/A	N/A	**74.2**
Flickr	Color	63.5	64.3	0.317	70.5
Flickr	Edge	65.6	65.1	0.376	69.7
Flickr	Gist	68.8	69.2	0.425	68.0
Flickr	Text	N/A	N/A	N/A	49.4

The results in bold indicate the best performance according to the evaluation criteria listed in the columns

- The Geograph dataset outperforms the Flickr dataset except when using the color histogram features.
- The relative performance of the visual features depends on the dataset. For the Geograph dataset, the ordering is the same as for the image level classification with the gist features performing best overall followed by the edge histogram features. This ordering is reversed for the Flickr dataset.
- The text features perform better than the visual features for the Geograph dataset but much worse for the Flickr dataset.

When comparing the tile level to the image level classification, we conclude:

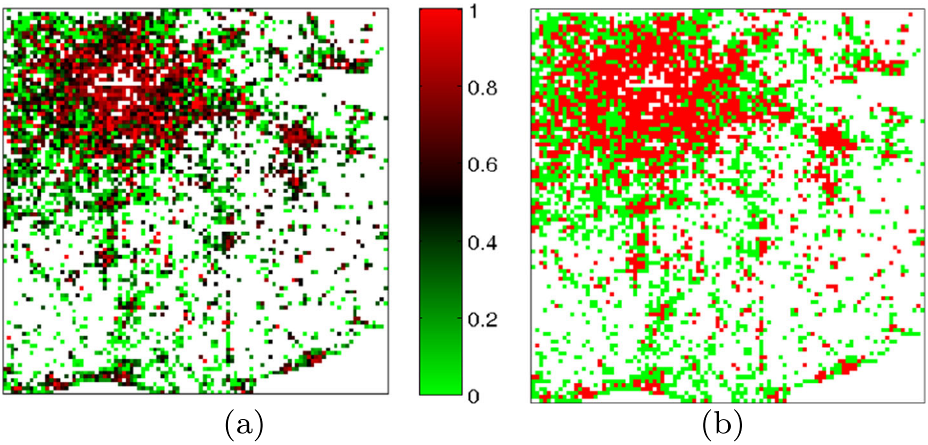


Fig. 7 The predicted fraction values **a** and binary labels **b** that result from using gist features to classify Geograph images as developed or undeveloped. Compare with the ground truth in Fig. 6

- Aggregating the color and edge histogram features at the tile level results in worse performance for the Geograph dataset but significantly better performance for the Flickr dataset.
- Aggregating the gist features at the tile level results in slightly reduced performance for both datasets.

3.4.3 Detailed results

Table 2 presents detailed results in the form of the true positive, true negative, false positive, and false negative counts and rates for each of the experimental configurations. We make the following conclusions based on these results:

- The image level SVM classifiers appear to over-classify the images as being developed. This bias is evident in the fact that the threshold applied to the fraction values in order for the the ratio of the developed to undeveloped tiles in the predicted set to match that of the ground truth is always greater than 0.5. The optimal threshold (not shown)

Table 2 Detailed classification results

Set	Lev	Fea	Th	TP	TN	FP	FN
Geo	Im	Col	Fix	1235 (74.6 %)	1632 (68.4 %)	754 (31.6 %)	420 (25.4 %)
Geo	Im	Col	Ad	1089 (65.8 %)	1860 (78.0 %)	526 (22.0 %)	566 (34.2 %)
Geo	Im	Edg	Fix	1310 (79.2 %)	1549 (64.9 %)	837 (35.1 %)	345 (20.8 %)
Geo	Im	Edg	Ad	1094 (66.1 %)	1859 (77.9 %)	527 (22.1 %)	561 (33.9 %)
Geo	Im	Gst	Fix	1228 (74.2 %)	1804 (75.6 %)	582 (24.4 %)	427 (25.8 %)
Geo	Im	Gst	Ad	1228 (74.2 %)	1804 (75.6 %)	582 (24.4 %)	427 (25.8 %)
Fli	Im	Col	Fix	1277 (77.2 %)	1291 (54.1 %)	1095 (45.9 %)	378 (22.8 %)
Fli	Im	Col	Ad	933 (56.4 %)	1664 (69.7 %)	722 (30.3 %)	722 (43.6 %)
Fli	Im	Edg	Fix	1358 (82.1 %)	1292 (54.1 %)	1094 (45.9 %)	297 (17.9 %)
Fli	Im	Edg	Ad	976 (59.0 %)	1656 (69.4 %)	730 (30.6 %)	679 (41.0 %)
Fli	Im	Gst	Fix	1258 (76.0 %)	1524 (63.9 %)	862 (36.1 %)	397 (24.0 %)
Fli	Im	Gst	Ad	1031 (62.3 %)	1764 (73.9 %)	622 (26.1 %)	624 (37.7 %)
Geo	Ti	Col	NA	1305 (78.9 %)	1475 (61.8 %)	911 (38.2 %)	350 (21.1 %)
Geo	Ti	Edg	NA	1302 (78.7 %)	1614 (67.6 %)	772 (32.4 %)	353 (21.3 %)
Geo	Ti	Gst	NA	1208 (73.0 %)	1784 (74.7 %)	602 (25.2 %)	447 (27.0 %)
Geo	Ti	Txt	NA	1061 (64.1 %)	1936 (81.1 %)	450 (18.9 %)	594 (35.9 %)
Fli	Ti	Col	NA	1055 (63.7 %)	1794 (75.2 %)	592 (24.5 %)	600 (36.3 %)
Fli	Ti	Edg	NA	1026 (62.0 %)	1792 (75.1 %)	594 (24.9 %)	629 (28.0 %)
Fli	Ti	Gst	NA	1228 (74.2 %)	1518 (63.6 %)	868 (36.4 %)	427 (25.8 %)
Fli	Ti	Txt	NA	1012 (61.1 %)	986 (41.3 %)	1400 (58.7 %)	643 (38.9 %)

The first column indicates the dataset. The second column indicates whether the classification is performed at the image or tile level. The third column indicates the feature. The fourth column indicates whether a fixed or adaptive threshold is used to derive the binary label from the fraction value. Columns five through eight indicate the true positive, true negative, false positive, and false negative rates in terms of the number of tiles and the percentage. The test dataset contains 4,041 tiles of which 1,655 have positive labels (labeled as developed in the ground truth) ranges from 0.51 for the Geograph-color case to 0.66 for the Flickr-edge case. The one exception is the Geograph-gist case for which a threshold of 0.5 is optimal.

- The edge histogram features result in the highest image level over-classification, followed by the color histogram features.
- The tile level SVM classifiers have mixed biases. With regard to the visual features, the Geograph-color, Geograph-edge, and Flickr-gist appear to be biased towards the developed class in that they have higher false positive than false negative rates. The Geograph-gist, Flickr-color, and Flickr-edge are biased towards the undeveloped class. The text features are heavily biased towards the undeveloped class for the Geograph dataset but heavily biased towards the developed class for the Flickr dataset.

3.5 Discussion

The experiments above demonstrate that large collections of geo-referenced ground level photos can be used to derive maps of what-is-where on the surface of the Earth. Binary land cover maps produced in an automated fashion using the visual and text features of the images were shown to be qualitatively and quantitatively similar to the ground truth. We now discuss further insights into the proposed framework.

We showed that image level classifiers could be learned in a weakly supervised manner by propagating the ground truth tile level labels to individual images located in those tiles. This has clear benefits in terms of the manual effort required to train the classifiers. While our training and test datasets are disjoint, they come from the same 100 × 100 km region. We plan to explore how well classifiers trained on one region generalize to other regions especially for which land cover maps are not available or are out-of-date.

The Geograph dataset outperforms the Flickr dataset confirming that photographer intent is important for treating community contributed photos as VGI. This is not an unexpected since the Geograph images are contributed by users whose goal is to collect geographically representative photographs. Since non-specialized collections such as Flickr have better coverage, this finding raises the interesting research question of how to use one dataset to improve another. Specifically, in what ways can the Geograph dataset be used to derive filters or other mechanisms for improving the Flickr dataset.

The gist features were shown to outperform the color and edge histogram features for classifying individual images as developed or undeveloped. This is in agreement with other studies on using gist features for scene classification. The visual features are complementary so combining them should result in improved performance.

While aggregating gist features at the tile level proved to be ineffective, tile level color and edge histograms were surprisingly effective especially for the Flickr dataset in which the tile level labeling significantly outperforms the image level labeling. This is interesting because it indicates that the aggregation helps remove the noise in non-specialized datasets such as Flickr. We plan to explore richer representations of the aggregate features such as Gaussian mixtures or non-parametric kernel density models.

The text based analysis was shown to be effective for the Geograph but not the Flickr dataset. This shows that photographer intent, here in terms of how individual images are described and tagged, seems to be even more of a factor when considering the text annotations than the visual content. Put differently, the fact that the visual aspects of the photo collections appear to be less affected by contributor intent suggests that it is preferable over text annotation as a source of VGI.

4 Removal of geographically uninformative images

Our results above show that the Flickr dataset performs worse than the Geograph dataset for land cover classification. This makes sense because the Flickr images were not necessarily contributed with the goal of documenting what-is-where on the surface of the Earth. We therefore investigate whether we can improve the performance by removing images that are unlikely to be geographically informative. We do this by exploiting the physical properties of the cameras as recorded in the image metadata, namely the type of camera used to capture the images and whether the flash fired.

We first consider whether an image was acquired with a stand-alone camera or with a mobile phone. The motivation here is that images from stand-alone cameras will generally be of better quality and thus more suitable for automated image analysis. We thus remove images that were acquired with mobile phones from the training data.

We next consider whether the flash fired when the image was taken. The motivation here is that the flash is a strong indication that the image was taken indoors and we observed that Flickr images taken indoors are generally not informative for land cover classification since there are a large number taken inside isolated buildings in undeveloped regions. Further, images taken outdoors with a flash usually capture close-by objects under low-light and thus are also not informative. We therefore remove images that were acquired with flash from the training data.

4.1 Experiments

To extract the camera properties, we use the Flickr API to obtain the EXIF metadata of each photo in the training set. We construct the *Camera* training set where photos taken by cameras in mobile phones are removed based on the camera models listed in the metadata. We also construct the *Flash off* training set in a similar fashion.

We follow the same experimental setup described in Section 3.3 except that the training data has now been filtered to remove geographically uninformative images. We use edge histograms as features but expect the results for the other features to be similar. The trained classifiers are applied to the whole test set–we do not filter the test images–so as to simulate the situation in which EXIF data might not be available for the set of target images.

Table 3 compares the performance of the different training sets. "Flickr" is the original, unfiltered training set; "Flickr Camera" is the training set with mobile phone images removed; and "Flickr Flash Off" is the training set with flash images removed. These results demonstrate that the filtered training sets result in slight performance gains. To further validate our hypothesis on camera flash, we also created a "Flickr Flash On" training set which contains only those images acquired with flash. As predicted, this results in worse performance.

Table 3 Image level classification results on Flickr dataset using edge histogram features

Training set	Overall class. rate (%)
Flickr	65.6
Flickr Camera	65.7
Flickr Flash Off	66.1
Flickr Flash On	64.8

Although the performance gains are modest, this experiment demonstrates that removing geographically uninformative images can improve land cover classification. Further work is needed, though, on exploiting the image metadata to improve this and other applications of proximate sensing.

5 Future research directions

The sections above presented a case study on using geo-referenced ground level photos for binary land cover classification. There are many interesting ways in which the proposed proximate sensing framework can be extended to other problems.

5.1 Algorithmic extensions

There are some algorithmic extensions which could improve the performance of the land cover classification above. The three different image features are complementary so combining them either pre-classifier through (possibly weighted) feature concatenation or post-classifier through classifier output fusion should improve the performance. Similar with the text features. Combining both the Flickr and Geograph images would also likely improve the performance. Also, since the map values in this case are real-valued, the developed fraction of a tile, it would be interesting to use a regression framework instead of classification.

5.2 Enhanced land cover and land use classification

An obvious extension is to increase the number of land cover classes beyond developed and undeveloped. Figure 2 lists the 10 aggregate classes in the LCM2000 of the United Kingdom. The United States Geological Survey (USGS) National Land Cover Database contains the following classes: open water; developed, open space; developed, low intensity; developed, medium intensity; developed, high intensity; barren land (rock/sand/clay); deciduous forest; evergreen forest; mixed forest; shrub/scrub; grassland/herbaceous; pasture/hay; cultivated crops; woody wetlands; and emergent herbaceous wetlands. Distinguishing between these classes will almost certainly require a richer set of visual features than the three used in the experiments above.

A particularly interesting aspect of ground level images is their potential for discriminating between land use classes. While remote sensing is useful for deriving maps of land cover which refers to the vegetation, structures, or other features that cover the land, it is much less effective at deriving maps of land use which refers instead to how the land is used by humans. Land parcels with different land uses, for example a hospital and a shopping center, might share similar land cover (building, parking lot) and thus be difficult to distinguish in overhead imagery. Proximate sensing instead relies on ground level images of objects and activities and thus could resolve such ambiguities.

The Standard Land Use Coding Manual [26] of the Urban Renewal Administration in the US Department of Commerce defines the following eight top-level land use classes: residential; manufacturing; transportation, communications, and utilities; trade; services; cultural, entertainment, and recreational; resource production and extraction; and undeveloped land and water areas. While some of these classes might be recognizable using remote sensed imagery, their subclasses are much more difficult. Trade is partitioned into several subclasses including building materials, hardware, and farm equipment; food; automotive; apparel and accessories; furniture; and eating and drinking. Services is partitioned into finance, insurance, and real estate; personal; repair; professional (which is further partitioned into medical, dental, etc.); governmental; and educational.

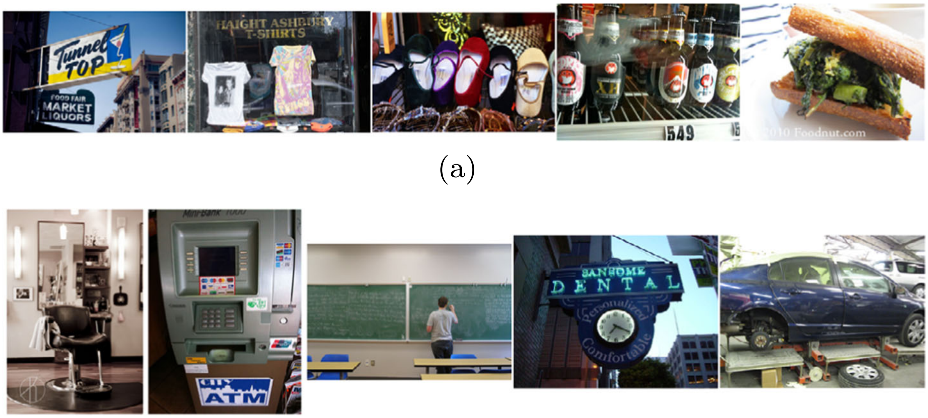
To demonstrate how proximate sensing might be useful for identifying these subclasses, we performed a location-constrained search in the downtown San Francisco area using the Flickr portal. Figure 8a shows five images sampled from the over 6,700 images identified using the keyword "shop". Clearly these images contain information about the trade subclasses. Figure 8b shows five images identified using several keywords ("barber", "ATM", "classroom", "dentist", and "car repair"). These images contain information about the services subclasses.

It is unlikely though that scene recognition based on global image features alone will be effective for land use classification especially for discriminating between subclasses. Object recognition techniques will likely be necessary. The problem of object recognition for geographic inference provides an interesting context in which to apply and evaluate existing techniques as well as develop novel ones. Significant progress has been made in object recognition over the past decade due to advances in image representations such as local invariant features [19], object modeling such as probabilistic generative models [27], and the availability of standardized test datasets such as Caltech-256 [9] which contains over 30,000 images of 256 object classes. These techniques and resources provide an excellent starting point to investigate proximate sensing for land use classification.

5.3 Analyzing georeferenced videos

Community contributed videos are another form of on-line media which could be useful for mapping what-is-where through proximate sensing. Flickr, for example, contains a growing collection of short videos (limited to 90 seconds) many of which have location information.

The problem of video analysis for geographic inference, like object recognition, provides an interesting context in which to investigate recent advances in event and activity recognition as well as develop new techniques. Videos have great potential for land use



(a)

(b)

Fig. 8 Georeferenced ground level photos could be used to perform land use classification. **a** Five images from Flickr which could be used to distinguish trade subclasses. **b** Five images which could be used to distinguish services subclasses

classification since many of the subclasses are associated with specific activities, such as the eating and drinking subclass of trade, or the different subclasses of the cultural, entertainment, and recreational top-level class.

5.4 Integrating spatial models

Tobler's first law of geography states that all things are related, but nearby things are more related than distant things [28]. Our work on proximate sensing has so far not considered this or other spatial models which have proven useful for remote sensing. Prior knowledge of the spatial distribution of developed and undeveloped regions could improve the binary land cover classification results above. And, spatial models will become even more important as the number of classes increases as well as for the more complex problem of land use classification. There is a wealth of spatial models which could be incorporated into the proximate sensing framework ranging from linear estimation like kriging to generative probabilistic models based on Markov random fields. However, it is not clear whether these models will spatially scale-down to the granularity of the analysis that is made possible by geo-referenced on-line media, or whether new models are required. This could be an interesting research topic for the spatial analysis community.

6 Conclusion

In the 20 minutes or so it has taken you to read this paper, over two thousand geo-referenced photos have been uploaded to Flickr (on average). These photos are a valuable source of VGI and potentially contain timely geographic information. The challenge to the multimedia content analysis and related computer science research communities is how to extract geographic information from this data.

In this context, we presented a novel framework based on proximate sensing and showed how it could be used to perform binary land cover classification into developed and undeveloped regions based on the visual content and text annotations of geo-referenced ground level images. We feel this is a good start to the problem but that many interesting challenges and opportunities remain ahead.

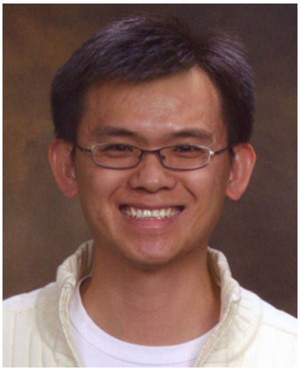
Acknowledgments This work was funded in part by a National Science Foundation CAREER grant (IIS-1150115) and a US Department of Energy Early Career Scientist and Engineer/PECASE award. We thank the anonymous reviewers for their informative feedback.

References

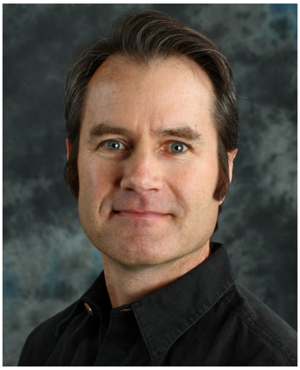
1. Cao L, Luo J, Kautz H, Huang T (2008) Annotating collections of photos using hierarchical event and scene models. In: Proceedings of the IEEE international conference on computer vision and pattern recognition, pp 1–8
2. Cao L, Yu J, Luo J, Huang TS (2009) Enhancing semantic and geographic annotation of Web images via logistic canonical correlation regression. In: Proceedings of the ACM international conference on multimedia, pp 125–134
3. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm
4. Chen WC, Battestini A, Gelfand N, Setlur V (2009) Visual summaries of popular landmarks from community photo collections. In: Proceedings of the ACM international conference on multimedia, pp 789–792

5. Crandall D, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the world's photos. In: Proceedings of the international world wide web conference, pp 761–770
6. Cristani M, Perina A, Castellani U, Murino V (2008) Geo-located image analysis using latent representations. In: Proceedings of the IEEE international conference on computer vision and pattern recognition, pp 1–8
7. Gallagher A, Joshi D, Yu J, Luo J (2009) Geo-location inference from image content and user tags. In: Proceedings of the IEEE international conference on computer vision and pattern recognition, workshop on internet vision, pp 55–62
8. Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211—221
9. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology
10. Hays J, Efros A (2008) IM2GPS: estimating geographic information from a single image. In: Proceedings of the IEEE international conference on computer vision and pattern recognition, pp 1–8
11. Hofmann T (1999) Probabilistic latent semantic indexing. In: SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pp 50–57
12. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42(1–2):177–196
13. Jacobs N, Satkin S, Roman N, Speyer R, Pless R (2007) Geolocating static cameras. In: Proceedings of the IEEE international conference on computer vision, pp 1–6
14. Joshi D, Luo J (2008) Inferring generic activities and events from image content and bags of geo-tags. In: Proceedings of the international conference on content-based image and video retrieval, pp 37–46
15. Kennedy L, Naaman M, Ahern S, Nair R, Rattenbury T (2007) How Flickr helps us make sense of the world: context and content in community-contributed media collections. In: Proceedings of the ACM international conference on multimedia, pp 631–640
16. Kennedy L, Naaman M (2008) Generating diverse and representative image search results for landmarks. In: Proceedings of the international world wide web conference, pp 297–306
17. Leung D, Newsam S (2009) Proximate sensing using georeferenced community contributed photo collections. In: ACM international conference on advances in geographic information systems: workshop on location based social networks
18. Leung D, Newsam S (2010) Proximate sensing: inferring what-is-where from georeferenced photo collections. In: Proceedings of the IEEE international conference on computer vision and pattern recognition
19. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
20. Manjunath BS, Ohm JR, Vasudevan VV, Yamada A (1998) Color and texture descriptors. IEEE Trans Circ Syst Video Technol 11:703–715
21. Moxley E, Kleban J, Manjunath BS (2008) SpiritTagger: a geo-aware tag suggestion tool mined from Flickr. In: Proceedings of the ACM international conference on multimedia information retrieval, pp 24–30
22. Naaman M, Yeh RB, Garcia-Molina H, Paepcke A (2005) Leveraging context to resolve identity in photo albums. In: Proceedings of the ACM/IEEE-CS joint conference on digital libraries, pp 178–187
23. Newsam S (2010) Community-contributed photo collections as volunteered geographic information: crowdsourcing what-is-where. IEEE Multimedia Spec Issue Knowl Discov Over Commun-Contributed Multimedia Data 17(4)
24. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
25. Quack T, Leibe B, Van Gool L (2008) World-scale mining of objects and events from community photo collections. In: Proceedings of the international conference on content-based image and video retrieval, pp 47–56
26. Standard Land Use Coding Manual. Urban Renewal Administration, Housing and Home Finance Agency and Bureau of Public Roads, Dept. of Commerce (1965)
27. Sivic J, Russell BC, Efros AA, Zisserman A, Freeman WT (2005) Discovering object categories in image collections. In: Proceedings of the IEEE international conference on computer vision
28. Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(2):234–240

29. Yanai K, Yaegashi K, Qiu B (2009) Detecting cultural differences using consumer-generated geotagged photos. In: Proceedings of the international workshop on location and the web
30. Zheng YT, Zhao M, Song Y, Adam H, Buddemeier U, Bissacco A, Brucher F, Chua TS, Neven H (2009) Tour the world: building a web-scale landmark recognition engine. In: Proceedings of the IEEE international conference on computer vision and pattern recognition, pp 1085–1092



Dr. Leung is a lecturer of the School of Engineering at the University of California at Merced. He received his Ph.D. from the University of California at Merced, his M.S. from the California State University at Fresno, and his B.S. from the University of Wisconsin-Madison. His research interests include computer vision, image processing, and geographic information system.



Dr. Newsam is an Associate Professor and Founding Faculty of Electrical Engineering and Computer Science at the University of California at Merced. He received his Ph.D. from the University of California at Santa Barbara, his M.S. from the University of California at Davis, and his B.S. from the University of California at Berkeley. Prior to joining UC Merced, he was a post-doctoral researcher at Lawrence Livermore National Laboratory. Dr. Newsam is the recipient of a U.S. Department of Energy Early Career Scientist and Engineer Award, a U.S. National Science Foundation Faculty Early Career Development (CAREER) Award, and a U.S. Office of Science and Technology Policy Presidential Early Career Award for Scientists and Engineers (PECASE). He is co-director of the Spatial Analysis and Research Center (SpARC) at UC Merced. His research interests include image processing, computer vision, pattern recognition, and data mining particularly as applied to scientific data.